Mining SNPs and linkage analysis in Cynara cardunculus

UNIVERSITÀ DEGLI STUDI DI TORINO

Mining SNPs and linkage analysis in *Cynara cardunculus*

Sergio Lanteri, Alberto Acquadro, Davide Scaglione, Ezio Portis

University of Torino, DIVAPRA, Via Leonardo da Vinci 44 – 10095 - Grugliasco (TO) – Italy; sergio.lanteri@unito.it

Abstract

Cynara cardunculus L., a member of the *Asteraceae* family, is a diploid (2n=34) outcrossing perennial species native to the Mediterranean basin. It includes globe artichoke (var *scolymus* L.), which today is grown as vegetable all over the world, cultivated cardoon (var. *altilis* DC), locally grown in Southern European countries, and their progenitor wild cardoon (var. *sylvestris* Lam). The species is also a valuable source of pharmaceutical compounds, and is exploitable for the production of lignocellulosic biomass as well as oil from seed, the latter being suitable for both edible and bio-fuel end-uses.

By crossing a non spiny globe artichoke genotype (female parent) with selected genotypes of the tree botanical taxa, we generated three F1 segregating progenies from which genetic maps, based on the two-way pseudo test cross strategy, have been developed. From the globe artichoke and cultivated cardoon genetic maps a reference SSR-based consensus map was constructed, which consists of 227 loci (217 SSRs and ten SNPs) assembled into 17 major linkage groups. To further saturate the *C. cardunculus* maps we recently applied NGS (next generation sequencing) technologies for mining a wide set of SNPs (single nucleotide polymorphism). Based on Illumina sequencing of gDNA RAD (restriction associated DNA) tags of three mapping parents (e.g. non spiny globe artichoke, cultivated and wild cardoon), we generated ~19.000 genomic contigs (mean 312 bp) and ~17.000 SNPs (density 1/139 bp). Side by side, the transcriptome of the same mapping parents was sequenced by using a 454 platform, and raw data *de novo* assembled and annotated to generate the first reference transcriptome of the species (38,726 unigenes, 32.7 Mbp).

The 454 reads, together with Illumina paired ends (PEs) from further eight *C. cardunculus* genotypes were aligned on the reference contig set, and ~195.000 SNPs were called (density 1/169bp in coding regions). The two workflows led to produce a massive set of SNPs in *C. cardunculus*, and made possible create an extensive gene catalogue as a valuable resource for upcoming genomic and genetic studies.

1.1 *The* Cynara cardunculus *complex*

Cynara cardunculus L. is native to the Mediterranean Basin and includes three botanical taxa: the globe artichoke (var. *scolymus*), the cultivated cardoon (var. *altilis*) and the wild cardoon [var. *sylvestris* (Lamk) Fiori]. The three forms are fully cross-compatible with one another, and form fertile hybrids (Basnizki and Zohary 1994). Reproductive barriers separate the *C. cardunculus* complex from the other *Cynara* species (Rottenberg et al. 1996). The crosses between *C. cardunculus* and the wild species *C. syriaca*, *C. algarbiensis*, *C. baetica* and *C. humilis* do all produce few seeds, although the hybrids are generally sterile; the wild species are therefore regarded as members of the secondary wild gene pool of *C. cardunculus* (Rottenberg and Zohary 2005). On both morphological (Wiklund 1992) and cytogenetic (Rottenberg et al. 1996) grounds, the closest of the wild species to the cultivated complex is *C. syriaca*. The monophyly and evolution of the *Cynara* spp. have been investigated by sequence comparisons between various ITS (internal transcribed spacer) and ETS (external transcribed spacer) regions (Robba et al. 2005; Sonnante et al. 2007) leading to the suggestion that the *cardunculus* complex is more differentiated and evolved than the other wild species.

Molecular (Lanteri et al. 2004; Acquadro et al. 2005), cytogenetic and isozyme (Rottenberg et al. 1996) studies have confirmed that wild cardoon is the ancestor of both the domesticate globe artichoke and cultivated cardoon, which evolved independently under the influence of distinct anthropogenic selection criteria. The earliest report of the presence of *C. cardunculus* in Sicily and Greece dates back to Theophrastus (371–287 BCE), while in 77 CE, the Roman naturalist Pliny the Elder mentioned its use for medicinal purposes; however, little is known either of the process of domestication or the subsequent diversification of the two taxa. It has been assumed that before globe artichoke was selected, cardoon was cultivated for its fleshy stems and roots, which were considered a delicacy by the ancient Greeks and Romans (Portis et al. 2005a; Portis et al. 2005b). On the other hand, the best guess is that the globe artichoke was domesticated and transformed into the plant that we know today, most probably between 800 and 1500 CE in family or monastery gardens. Recently, by assessing the AFLP pattern of genetic diversity of a collection of Sicilian globe artichoke landraces, which have been cultivated for a number of centuries by local farmers, one landrace was identified which appears to represent an early stage of the domestication process, suggesting Sicily as one of the possible centre of globe artichoke domestication (Mauro et al. 2008).

Globe artichoke contributes significantly to the Mediterranean agricultural economy, with an annual production of about 750 metric tons (MT) from over 80,000 ha of cultivated land and with an annual turnover exceeding US $ 500 million. Italy is the leading world producer (480 MT/year, FAOSTAT 2010), followed by Egypt and Spain. Globe artichoke cultivation is increasing in South

America and the United States, and more recently also in China. The prime globe artichoke product consists of the immature inflorescence (heads of capitula), which can be consumed in fresh, canned or frozen form. Each plant produces a number of capitula, the largest of which (the main capitulum) merges from the apex of the central stem, while the smaller ones are produced on lateral branches.

Italy has the richest globe artichoke primary cultivated "gene pool" and harbours many distinct clonal varietal groups, best adapted to local environments. On the basis of harvest time, varietal types can be classified as early, producing heads from autumn to spring, and late, producing heads from early to late spring. On the basis of capitulum characters, cultivated germplasm has been classified into four main groups: (1) the Spinosi group, containing types with long sharp spines on bracts and leaves; (2) the Violetti group, with medium-sized, violet-coloured and less spiny heads; (3) the Romaneschi group, with spherical or subspherical non-spiny heads; (4) the Catanesi group, with relatively small, elongated and non-spiny heads. The classification based on head is in consistent agreement with the one obtained by assessing the AFLP genetic variation in a wide collection of 84 varietal types grown worldwide, indicating that the cultivated morphotypes play an important role in determining variation within the cultivated globe artichoke germplasm (Lanteri et al. 2004). Although in recent years some seed (achenes)-propagated varieties have been introduced, but vegetative propagation, by means of basal and lateral offshoots (either semi-dormant or actively growing), or stump pieces, has been adopted for centuries, and it is still largely prevalent in most of the varietal types and local landraces. Due to the limited selection adopted by farmers on the mother plants used for vegetative propagation, as well as mutations occurred over time, the populations at present in cultivation are multiclonal and characterized by a wide range of within population genetic variation (Lanteri et al. 2001; Portis et al. 2005c).

The cultivated cardoon (*C. cardunculus* var. *altilis* DC) is usually raised from seed and handled as annual crop; its cultivation is much less widespread than that of the globe artichoke and the crop remains of regional importance in Spain, Italy and the south of France, where it is used in traditional dishes. The edible parts of the plant are the fleshy stems which are typically collected in late autumn-early winter and often, before collection are tied together, wrapped in straw, and/or buried for about three weeks in order to accentuate the flavour. A study based on SSR and AFLP profiling of the most widely grown Italian and Spanish local varieties showed that they form two separate gene-pools and that a considerable level of within variety variation is present (Portis et al. 2005b).

The wild cardoon is a robust thistle distributed over the west and central part of the Mediterranean basin (Portugal to west Turkey) as well as Canary Islands; in post Columbian time it colonized some part of the New World and has spread as a weed in Argentina and California

(Marushia and Holt 2006). Its flowers have been used for centuries in the Iberian Peninsula for manufacturing of ovine and caprine milk cheese (Sousa and Malcata 1996; Barbagallo et al. 2007) and its small and thorny capitula are sometimes sold in local markets in Sicily (Ierna and Mauromicale 2010).

1.2 Uses of globe artichoke and cardoon other than for human food

C. cardunculus has long been known to represent a valuable source of biopharmaceutical compounds (Slanina et al. 1993; Wagenbreth 1996; Sevcikova et al. 2002; Wang et al. 2003). Roots and rhizomes, used also for brew or infusion, provide a source of inulin, a demonstrated enhancer of the human intestinal flora, while leaves and heads represent one of the richest natural source of compounds originating from the metabolism of phenylpropanoids, with caffeoylquinic acids and flavonoids as major components. *C. cardunculus* extracts influence glucose and lipid metabolism (Blumenthal et al. 2000) and were reported to be effective in increasing the feeling of satiety in overweight subjects (Rondanelli et al. 2011); in various pharmacological test systems it has been demonstrated that they (i) protect proteins lipids and DNA from oxidative damage from free radicals (Gebhardt 1997; Brown and Rice-Evans 1998; Perez-Garcia et al. 2000), (ii) inhibit cholesterol biosynthesis and contribute to the prevention of atherosclerosis and other vascular disorders (Kraft 1997; Brown and Rice-Evans 1998; Gebhardt 1998; Pittlern and Ernst 1998; Matsui et al. 2006; Bundy et al. 2008). Furthermore, it has been demonstrated that *C. cardunculus* extracts inhibit HIV integrase, a key player in HIV replication and its insertion into host DNA (McDougall et al. 1998; Slanina et al. 2001), possess apoptotic properties (Miccadei et al. 2008) and exert antibacterial activity (Martino et al. 1999).

The composition of the globe artichoke phenolic fraction includes four mono-caffeoylquinic isomers, six dicaffeoylquinic isomers, six flavonoid glycosides, and at least seven anthocyanins (Lattanzio et al. 2009). The genes involved in the biosynthesis of the mono-caffeoylquinic acid (chlorogenic acid) have been identified as well as their regulation under UV-C stress (Comino et al. 2007, 2009; De Paolis et al. 2008; Moglia et al. 2009; Menin et al. 2010; Sonnante et al. 2010). Conversely, the biosynthetic pathway leading to di-caffeoylquinic acids is a matter of debate (Villegas and Kojima 1986; Hoffmann et al. 2003; Niggeweg et al. 2004).

The characteristic bitterness of both globe artichoke and cultivated cardoon is mainly due to the presence of sesquiterpene lactones (STLs), of which the two major representatives are cynaropicrin and, at lower concentration, grosheimin and its derivatives (Schneider and Thiele 1974; Cravotto et al. 2005). Cynaropicrin, like many sesquiterpenes lactones, has various medicinal activities (Shimoda et al. 2003; Cho et al. 2004; Schinor et al. 2004; Emendorfer et al. 2005; Ishida

et al. 2010) among which cytotoxicity against several types of cancer cells (Yasukawa et al. 2010). In globe artichoke a germacrene A synthase, involved in the first step of STLs biosynthesis, has been recently isolated, functionally characterized and mapped (Menin et al. 2012).

C. cardunculus has great potential as a source of renewable energy, thanks to its productivity of lignocellulosic biomass. The caloric value of the three *C. cardunculus* taxa is analogous, however cultivated cardoon has the highest biomass yield, which can reach up to ~ 19t/ha dry matter with an energy value ~ 17 MJ/kg (Angelini et al. 2009; Ierna and Mauromicale 2010; Portis et al. 2010; Ierna et al. 2012). The species has been also identified as a candidate for the production of seed oil which is suitable for both comestible and bio-fuel end-uses. Seed yield in cardoon is about 2 t/ha and up to 0.8 t/ha in globe artichoke (at 5% w/v moisture), from about 25 to 30% of which is oil of good alimentary quality (Foti et al. 1999) due to its high and well balanced content of oleic and linoleic acids, its low content of free acids, peroxides, saturated and linoleic acids and a favourable α-tocopherol content (Maccarone et al. 1999), while the seed material left after oil extraction can be used as a component of animal feed.

1.3 Linkage analyses: state of the art

The genome organization of *C. cardunculus* (2*n*=2*x*=34; haploid genome size ~1.08 Gbp), unlike other species belonging to the Asteraceae family (e.g. sunflower, lettuce and chicory), remains largely unexplored. The species is an out-breeder, and is characteristically highly heterozygous. Its marked level of inbreeding depression inhibits the use of backcross, F_2 or recombinant inbred populations for mapping purposes. As haploid induction - via either androgenesis or gynogenesis - has not yet been achieved (Motzo and Deidda 1993; Chatelet et al. 2005; Stamigna et al. 2005), no possibility is presently available to generate doubled haploid populations. Thus, genetic mapping in the species has relied on a double pseudo-testcross approach (Grattapaglia and Sederoff 1994), in which segregating F_1 progeny are derived from a cross between two heterozygous individuals.

The first genetic maps of *C. cardunculus* were provided by Lanteri et al. (2006), and based on a cross between two genotypes of globe artichoke, namely the varietal types ‘Romanesco C3’ (a late-maturing non-spiny type used as female) and ‘Spinoso di Palermo’ (an early-maturing spiny type used as male). This population was genotyped using a number of PCR-based marker platforms, resulting in a ~1300 cM female map consisting of 204 loci, divided into 18 linkage groups (LGs), and a ~1200 cM male map comprising 180 loci and 17 LGs. The two maps shared 78 loci, which allowed for the alignment of 16 of the LGs. The maps have since been extended by the inclusion of three genes involved in the synthesis of caffeoylquinic acid (Comino et al. 2009; Moglia et al. 2009)

and a number of microsatellite loci, of which 19 were represented in both maps (Acquadro et al. 2009).

New maps have lately been generated from a set of F_1 progeny involving the cross between the same female parent as previously ('Romanesco C3') and the cultivated cardoon genotype 'Altilis 41' (Portis et al. 2009a). The cultivated cardoon map comprised 177 loci, subdivided into 17 LGs and spanning just over 1000 cM, while the globe artichoke one featured 326 loci arranged into 20 LGs, spanning ~1500 cM with a mean inter-marker distance of ~ 4.5 cM. A set of 84 loci shared between this 'Romanesco C3' map and the previously developed one (Lanteri et al. 2006) allowed for map alignment and the definition of 17 homologous LGs, corresponding to the haploid chromosome number of the species. Later on, the maps have been integrated with the inclusion of all the genes involved in the synthesis of caffeoylquinic acids known in the species (Menin et al. 2010).

Since more markers were needed to saturate the maps, a further wide set of SSR markers was developed from ESTs (expressed sequence tags) of globe artichoke, made available by the Composite Genome Project (CGP; http://compgenomics.ucdavis.edu/). Using a custom bioinformatic pipeline, 36,321 ESTs were assembled into 19,055 unigenes (6,621 contigs and 12,434 singletons), annotated, and mined for perfect SSRs. Over 4,000 potential EST-SSR loci, lying within some 3,300 genes (one SSR per 3.6 kbp) have been identified, and PCR primers for the amplification of more than 2,000 of these have been designed. In a test of a sample of 300 of these assays, over half proved to be informative between the parents of the available mapping populations (Scaglione et al. 2009). As a result, a large number of these EST-SSR loci have been integrated into the globe artichoke and cultivated cardoon maps (Portis et al. 2012) and, more recently, exploited for genetic mapping in a population obtained by crossing globe artichoke with wild cardoon (Sonnante et al. 2011). The integration of 139 EST-SSR loci has significantly improved the resolution and accuracy of the 'Romanesco C3' and 'Altilis 41' maps. The female map was built with 473 loci spanning 1.544 cM with a mean inter-marker distance of 3.4 cM, corresponding to a 3.8% increase in length over the earlier map, but in a ~ 28% decrease in the mean inter-marker distance. The male map consisted of 273 loci spanning 1486 cM, with a mean inter-marker distance of 5.4 cM, representing a marked increase in both length (+42%) and number of loci (+50%), together with a minor decrease in the mean inter-marker distance (-5%). The two maps shared 66 codominant loci (64 SSRs and two SNPs), which allowed for the alignment of all the 'Romanesco C3' with the 'Altilis 41' LGs. Following alignment a consensus linkage map based exclusively on microsatellite and SNP markers (as depicted in Figure 1) was constructed (Portis et al. 2009b). The consensus map is shown in Figure 2; it comprised 227 loci (217 SSRs and ten SNPs targeting genes

involved in the synthesis of caffeoylquinic acids) arranged into 20 LGs (LOD threshold > 6.0). The consensus map length was 1068.0 cM, with a mean inter-marker spacing of 5.2 cM . The length of LGs varied from 4.0 to 113.7 cM (mean 62.8 cM), with the largest LG containing 36 loci. Lowering the LOD threshold to 5.0 resulted in the merging of three pairs of LG, thereby reducing the overall number to 17, corresponding to the haploid chromosome complement of the species. The majority of the LGs contained a mixture of 'Romanesco C3', 'Altilis 41' and shared co-dominant markers, with only four (LG_9, 13, 14 and 7) carrying shared loci and markers only present in the 'Romanesco C3' map.

Putative functions have been deduced for SSR markers derived from ESTs using homology searches with public protein databases. Annotation of mapped loci was performed via BlastX search as well as InterPro scan and GO categorisation made it possible to tag some biological functions. A set of 17 EST-SSR markers were annotated with GO terms involved in the 'response to stimulus' (Table 1), five and eight of which were derived from transcripts related to response to cold and salt stress, respectively. As an example, the marker CyEM-42, developed from the contig CL4773Contig1 (Scaglione et al. 2009) and mapped on LG_12 of the SSR-based consensus map, showed high aminoacidic similarity (81%) with the protein kinase PBS1 of *Arabidopsis*. To consider reliable orthology, a reciprocal tblastx analysis against the whole EST collection currently available for *C. cardunculus* was performed, and no better alignment than that of contig CL4773 was detected. PBS1 was found to work as R gene against the bacterial pathogen *Pseudomonas syringae*, where its cleavage, operated by the pathogens' effector AvrPphB, triggers the signalling cascade, generating the host response (Shao et al. 2002). *Pseudomonas* spp. together with other endophytic bacteria may affect globe artichoke plants both in field and during micropropagation (Penalver et al. 1994) and the CyEM-42 may be likely considered a reliable marker for tagging a bacterial resistance trait in the species. On the whole, these EST-SSR markers may be defined as functional markers with the potential to target polymorphisms in gene responsible for traits of interest; they can be also particularly useful for constructing comparative framework maps with other Asteraceae, giving the possibility to amplify ortholog genes and provide anchor loci.

This SSR-based consensus map of *C. cardunculus* is based on a robust marker platform of SSRs and a few gene-based SNP loci. It is expected that the further positioning of markers within target regions will provide key tools for marker-assisted breeding programs as well as the necessary framework to exploit mapping data obtained from diverse populations. At present, ~ 200 of the loci on the consensus map (about 88%) are sited within genic sequence, presenting some opportunity to identify candidate genes for particular traits within the species.

1.4 SNP mining

The first set of SNP (single nucleotide polymorphism) markers available for the species has been developed on genes involved in the synthesis of caffeoylquinic acid, as above reported. The allelic forms of globe artichoke acyltransferases HCT, HQT (Comino et al. 2007; Comino et al. 2009) and the hydroxylase C3'H (Moglia et al. 2009), were analysed in the two globe artichoke parental genotypes ('Romanesco C3' and 'Spinoso di Palermo') of the first genetic maps (Lanteri et al. 2006) and SNPs were identified. SNP genotyping of the F_1 progeny was carried out with the tetra-primers ARMS-PCR method (Ye et al. 2001; Chiapparino et al. 2004). A further SNP set has been later on developed by Menin et al. (2010) on three acyltransferases and on the *C4H*, *4CL* and *MYB12* genes, identified by an *in silico* scan of the globe artichoke unigene set assembled by Scaglione et al. (2009). Gene homologues were re-sequenced in the parental genotypes (globe artichoke 'Romanesco C3' and cultivated cardoon 'Altilis 41') of the genetic maps developed by Portis et al. (2009a) and genes successfully mapped.

Recent advances in next-generation DNA sequencing technologies have made possible the development of high-throughput SNP genotyping platforms, that allow for the simultaneous interrogation of thousands of SNPs. Such resources have the potential to facilitate the rapid development of high-density genetic maps, and to enable genome-wide association studies as well as molecular breeding approaches in a variety of taxa (Bachlava et al. 2012). Thousands of SNPs have been recently developed in *C. cardunculus* by Next-Generation Sequencing (NGS) technology using two complementary approaches (Figure 3):

1) genomic RAD (Restriction-site Associated DNA) tag sequencing (Miller et al. 2007) in combination with the Illumina Genome Analyzer sequencing device (Baird et al. 2008) of three genotypes (globe artichoke, cultivated cardoon and wild cardoon) that were crossed for developing F_1 mapping populations (Scaglione et al. 2012a);

2) transcriptome sequencing, via 454 and Illumina technologies, of the same three genotypes plus eight, five of which were globe artichoke, two cultivated and one wild cardoon (Scaglione et al. 2012b). Alongside, a functional characterisation and annotation of the obtained sequence set was performed. These SNPs represent a one-stop resource to produce a dense *C. cardunculus* genetic map via high-throughput genotyping technologies.

1.4.1 Genomic SNP mining

The recently developed restriction-site associated DNA (RAD) approach (Box 1) has been combined with the Illumina DNA sequencing platform to enable the rapid and mass discovery of SNP markers. Three genomic RAD libraries were obtained from the three *C. cardunculus* genotypes

belonging to the three taxa of the species and parents of two mapping populations. The first mapping population is an F_1 progeny involving the cross between globe artichoke ('Romanesco C3', female parent) and cultivated cardoon (genotype 'Altilis 41') (Portis et al. 2009a). The second one is an F_1 progeny involving the cross between the same female parent as previously and the wild cardoon (genotype 'Creta4') (Lanteri et al. 2011).

1.4.1.1 RAD tag sequencing and *de novo* assembly

The RAD-seq exercise produced 9.7 million reads (19.4 million Pair End - PE), equivalent to ~ 1 Gbp of sequence. The distribution of reads was uneven across the three DNA samples, with 1.2 million reads achieved for globe artichoke, 2.6 million for cultivated cardoon and 5.9 million for wild cardoon; the latter, being the largest set, was chosen as the basis for *de novo* contigs assembly. The assembly procedure created 19,061 reference genomic contigs, spanning 6.11 Mbp (with N50 = 321 bp and a mean a contig length of 312 bp). The GC content was ~ 37.4% which is similar to that of many dicots species (Jaillon et al. 2007) and represents the first survey on the base composition of the *C. cardunculus* genome.

1.4.1.2 RAD tag annotation

The contig sequences characterisation was conducted using the BlastX algorithm and it resulted in the annotation of 5,335 contigs (28.0%). Regardless of the genome-wide RAD sampling, a noteworthy part of the annotated sequences might be represented by coding regions, since a methylation-sensitive enzyme (*Pst*I) was used to produce the RAD-tag libraries (Palmer et al. 2003), although the rather short length of the RAD contigs made difficult to distinguish between putative genes and pseudogenes. Enzyme codes were retrieved for 1,327 contigs, defining a unique set of 313 putative enzymatic activities, which were mapped onto KEGG reference pathways (http://www.genome.jp/kegg/). The remaining portion of the contig set (72%) was not attributed to any known sequence, likely due to the RAD contigs shortness.

The transposable DNA element footprints detected, using RepeatMasker software (v3.2.9; http://www.repeatmasker.org) implemented with the RMBlast algorithm, and adopting the *Viridiplantae* repeats as reference, accounted for a 0.2% of the sequence, while 1.2% of the sequences derived from LTR retroelements, including Ty/Copia-like (0.8%) and Gypsy-like (0.2%). This quantification of transposable element abundance could have been underestimated, but these data represents a useful snapshot of relative abundance of each different mobile element class in *C. cardunculus*.

1.4.1.3 SNP calling

The PE sequences generated for each mapping parent were aligned using the reference contig set as a scaffold. In total, ~ 33,000 sequence variants were detected, including 1,520 short indels, distributed over 12,068 contigs. The overall SNP frequency was estimated to be 5.6 per 1,000 nucleotides, a level which is almost equal to that found in the non-coding regions of the *V. vinifera* genome (5.5 per 1,000 nucleotides; Velasco et al. 2007) and very similar to that observed in *Citrus* spp. ESTs (6.1 per 1,000 nucleotides; Jiang et al. 2010). A subset of ~ 17,400 SNPs was obtained considering allelic variant which were informative for both mapping populations (16,727 SNPs, and 723 1-2nt indels) distributed over 7,478 contigs.

Since *C. cardunculus* is highly heterozygous, SNPs were categorized as intra- or inter-varietal, where the former also represents the heterozygous state of the analysed genotype. The two types were not exclusive, therefore heterozygous SNPs present in one sample could be found in both heterozygous or homozygous states in other genotypes. The number of heterozygous SNP loci was 1,235 in the globe artichoke, 2,868 in the cultivated cardoon and 5,069 in the wild cardoon mapping parents (Figure 4). Heterozygous SNPs are of key importance for mapping studies since for the linkage analysis a two-way pseudo-testcross approach, based on a segregant F_1 progeny, was adopted. In this sense, a key parameter for the successful isolation of such useful SNP markers was the sequencing coverage.

1.4.2 Transcriptomic SNP mining

A total of eleven *C. cardunculus* EST libraries were produced and after the normalisation procedures, they were separately sequenced. Three libraries, deriving from the three mapping parents (Table 2), were sequenced with the 454 Titanium (Roche) to produce a reference transcriptome. Eight libraries, set up from five globe artichoke accessions, two cultivated cardoon and one wild cardoon genotypes (Table 3), were sequenced using the Illumina GAIIx platform, in order to highly increase the total SNP calling amount.

1.4.2.1 EST sequencing and *de novo* assembly

The outcome of 454-based cDNA sequencing of the three mapping parents generated some 1.7 M reads of overall length 695 Mb, which were reduced to 692 Mb after a post-sequencing filtering. The mean read length was equal to 392 bp (Table 2). cDNA libraries of other eight genotypes (Table 3) were sequenced using a GAIIx platform (Illumina) producing 6.9 Gbp of raw data (46.4 M paired-end reads) with a mean of 5.8 M reads per accession. The data set was reduced to 6.7 Gbp following the removal of adaptor sequences and other contaminants, and it was further reduced to

6.2 Gbp after quality trimming. For the *de novo* assembly process only the 454 reads were considered, while the Illumina data were simply adopted to increase the efficiency of the SNP mining process.

The assembly of 454 reads was achieved by a two-tier approach using the MIRA assembler ver.3.2.0 (Chevreux et al. 2004). In a first step, each individual sample was assembled independently. The process generated 37,622 contigs for 'Romanesco C3', 40,130 contigs for 'Altilis 41', and 42,837 contigs for 'Creta 4' with N50 contig lengths of 834 bp, 761 bp and 772 bp, and mean coverage levels of 7.31, 8.45 and 9.17X, respectively. For the 'Romanesco C3' assembly, a subset of 11,276 contigs resulted from the incorporation of a prior set of 28,641 Sanger ESTs (www.ncbi.nlm.nih.gov/dbEST). Then, after contaminant removal by BLASTX analysis, the three datasets were merged into a set of 38,726 contigs. This "reference" assembly spanned 32.7 Mbp and had a GC content of 42.1%. The mean contig length was 844.3 bp (N50: 951 bp).

A second assembly phase was carried out by merging at least two *taxon*-derived contigs from the first phase, and 20,469 contigs were generated. They consisted of a subset with a mean length of 1054 bp, while 5,375, 6,669 and 6,213 remained as single *taxon*-derived contigs of var. *scolymus*, var. *altilis* and var. *sylvestris*, respectively.

1.4.2.3 Sequence analysis and functional annotation

The sequence reads were assembled into 38,726 reference transcripts, which were successfully annotated, using the Blast2GO pipeline, by gene ontology terms via Blast and InterPro analyses. Enzymes were tagged on KEGG's reference pathways (www.genome.jp/kegg/), including primary and secondary metabolisms. On the whole, 16,419 enzyme codes were retrieved (12,449 transcripts) and subsequently mapped onto KEGG's pathways. The sample of *C. cardunculus* enzymes consisted of 1,133 unique enzyme codes distributed across 147 pathways. To provide an example, by analyzing the whole transcriptome complement, a subset of 71 enzymatic activities involved in phenylpropanoid synthesis were identified; 21 of these were annotated at varying levels of redundancy in the core phenylpropanoid pathway (KEGG's map: 00940), in which the synthesis of caffeoylquinic and di-caffeoylquinic acids (CQAs and dCQAs) takes place (Figure 5).

Transcriptional factor function was assigned to 1,398 transcripts, scattered across 67 families, while 316 sequences were tagged as candidate Resistance Gene Analogs (RGAs). Each sequence was scanned for the presence of recognition sites for known plant miRNAs. In total, target annealing sites for 302 miRNAs were located in 1,043 transcripts, which mainly belong to the

categories: “defense response” and “programmed cell death/apoptosis”, “reproduction”, “development of anatomical structure”, “photosynthesis”, “transmembrane receptor activity” and “transcription factor activity”. The 454-based assembly included non-nuclear transcripts. The *C. cardunculus* chloroplast genes identification was based on similarity to those of lettuce and sunflower (Timme et al. 2007) leading to the categorization of 137 contigs, of which 80 were putatively assigned the chloroplast genome. Similarly, the grapevine (*Vitis vinifera*) mitochondrial genes (Goremykin et al. 2009) aided in the identification of 52 *C. cardunculus* contigs, which were putatively attributed to the mitochondrial genome.

To estimate the transcriptome representation and its gene-level redundancy (e.g. splicing variants), two different approaches were adopted. Using the *A. thaliana* gene content, the 454 sequencing output was predicted to be assembled in a total of 29.3 Mbp, distributed in 24,064 unigenes (average length of 1,216 bp) which covered 96% of the transcriptome. Alternatively, the final contig set (38,000) was clustered by collapsing gene variants (e.g. alternative splicing), which generated a set of 29,830 unigenes that represents a *bona fide* estimation of the gene content of *C. cardunculus*. Data suggest that 23% of splicing variants could be present in the transcriptome assembly.

1.4.2.4 Read mapping and SNP calling

About 1.5 M of the 454-derived reads were aligned to the reference contig set (38,726 contigs). This number was reduced to ~ 1.0 M by removing those that showed more than one unique alignment, thereby lowering the risk of false SNP calls due to misalignment of paralog-derived reads or to redundancy resulting from splicing variants. The same procedure was repeated for the Illumina-derived reads, producing an alignment of ~ 60 M paired ends. Resolving paired ends reduced this to a set of ~ 21 M reads.

An assembly based on about 35 M sequences was generated by merging the 454 and Illumina sequence datasets, resulting in a median reference transcriptome coverage of 96X with 26,990 reference contigs containing at least 20 mapped reads. Reliable SNPs (Bayesian probability >95%) were detected at 195,400 sites across the set of 11 accessions. The average SNP frequency was calculated at one per 167 bp, with a mean of five per contig. Each SNP site was interrogated by scoring for the presence of at least one accession-specific sequence. Sequence information was available from an average of nine accessions per SNP site, and a core subset of 57,125 SNPs showed coverage from all the samples. The merging of the Illumina-derived reads (eight

accessions) with 454-generated reads substantially increased the number of parent-specific SNPs that were identified (Figure 6).

SNP frequency in the *C. cardunculus* transcriptome appears to be comparable to that found in the heterozygous grapevine whole genome sequence (Velasco et al. 2007) and among *Citrus* ssp. ESTs (Jiang et al. 2010). Overall, SNPs were most frequent in 3'-UTR (one per 126 bp), followed by the CDS (one per 169 bp), and the 5'-UTR (one per 265 bp). Within the UTRs, the frequency also matched that obtained in tomato expressed sequence (Jimenez-Gomez and Maloof 2009), while it was markedly different to that present in the coding region (~ 2 per kb). This discrepancy may reflect either the greater tolerance by the heterozygous state of non-synonymous substitutions, or merely is an ascertainment bias due to the analysis of a larger germplasm panel which also included accessions of a wild relative.

In *C. cardunculus*, as previously pointed out, the presence of intra-accession allelic variation is of particular interest. As expected by their shallower coverage, the 454-derived sequences produced a somewhat lower frequency of SNPs with successful heterozygous SNP calling (Figure 7). 'Altilis 41' was relatively the least heterozygous of the accessions (17,570 loci), as has been observed previously (Portis et al. 2005b; Portis et al. 2009a), while 'Romanesco Zorzi' was the most heterozygous (43,387 loci), followed by 'Violetto di Chioggia' (41,824 loci). 'Imperial Star' had the lowest ratio of heterozygous variants among globe artichoke genotypes (13.5%), which likely reflects its development from crosses among less genetically differentiated genotypes.

1.4.3 Conclusions

The second generation technologies provide high sequencing throughputs at significantly reduced costs if compared to Sanger. These platforms are currently employed for large-scale SNP discovery projects and, for medium-scale projects, they have been frequently applied in combination with reduced-complexity libraries, targeting genomic subsets.

One such method, aimed at decreasing the sample complexity, is to build up a genomic library, with a reduced locus representation including only a subset of sequences generated by restriction enzymes, which cut at frequent intervals throughout the genome. The generation of an SNP set can be achieved through the deep-sequencing of such libraries and the comparison between allelic variants can identify thousands of SNPs.

The recent RAD (Baird et al. 2008) approach is focused on the targeting of a discrete number of genomic regions adjacent to specific restriction sites, and it can effectively reduce the number of the fragments to be sequenced in a given complex genome. This strategy (see Box 1) represents a promising experimental scheme in term of costs and technical simplicity and, so far,

441 has been successfully adopted for SNP discovery in many plant and animal species (Davey et al.
442 2011).

443 An alternative approach is to focus onto the transcriptome deep-sequencing, which reduces
444 the representation of low information-content repetitive sequences in species possessing a large
445 genome and/or without a finished genome sequencing project. An EST library can lead to identify a
446 large number of genetic loci and targeting SNPs in coding sequences. This kind of library represents
447 a one-stop resource useful for many downstream applications and to address many biological
448 questions in plant science. It can aid the identification of genes underlying phenotypes of interest
449 through the development of expression arrays or provide thousands of loci as a source of potential
450 markers for QTL mapping applications and population genomic studies.

451 The two experimental workflows led to produce a massive set of SNPs in *C. cardunculus*, and
452 made possible create an extensive gene catalogue, as a valuable resource for upcoming genomic and
453 genetic studies. Both approaches have proven to be efficient for SNP mining, although
454 characterized by peculiarities and limitations which deserve to be considered in view of specific
455 research targets. In *C. cardunculus* the EST sequencing approach generated a set of reference
456 coding sequences spanning 32.7 Mbps, establishing a 'general gene catalog' of 38,726 as *bona fide*
457 representation of the transcriptome. In contrast the RAD-tag sequencing approach permitted to
458 sequence 6.0 Mbps separated in lesser and shorter number of contigs (~ 19,000; 28% of which were
459 annotated as CDS-like). The number of SNPs was higher for EST than for RAD-tag approach
460 (195,000 *vs*. 34,000); nevertheless, the SNP frequency observed in the two pipelines were
461 somewhat comparable (5.9 *vs*. 5.6 per 1,000 nt). The RAD-tags data revealed to be extremely
462 informative to preliminary survey the repetitive DNA component of the *C. cardunculus* genome,
463 and allowed us to make some inferences regarding the contribution of DNA methylation in
464 inhibiting its expansion (Scaglione et al. 2012a).

465 From the standpoint of costs, RAD technology was attempted with a great technical
466 simplicity and a low cost/time expense. The cDNA library setting up was indeed more complex for
467 both the need of standardization/normalisation procedures and some extra enzymatic steps required,
468 however, side by side, its sequencing output provided a better picture of the globe artichoke coding
469 genome. Bearing in mind a future in which the globe artichoke genome will be completely
470 sequenced and publicly available, the genomic RAD approach may represent one of the most
471 feasible and cheap strategy for accomplishing affordable targeted re-sequencing projects. It is also
472 likely that the increasingly lowering of sequencing costs will make the scientific community to
473 converge towards new approaches of 'genotyping-by-sequencing'. This scheme proceeds to explore

all the nucleotidic positions of a genome in a single experiment, and will permit an integration of mapping and sequencing steps, likely bypassing many costly physical mapping procedures.

The combination of two NGS platforms (454 FLX Titanium - Roche and GAIIx - Illumina) for the extensive characterization of the genome and transcriptome of *C. cardunculus*, has proven to be a highly reliable tools for SNP discovery. Overall, the availability of such a large number of sequence-based markers, in a format allowing for high-throughput genotyping, offers opportunities to developed a high-density genetic map and association mapping studies aimed at correlating molecular polymorphisms with variation in phenotypic traits, as well as for molecular breeding approaches in a species which has multiple end-uses such as food, nutraceuticals and bioenergy. The high number of mined SNPs represents also an excellent resource for evolutionary genetic studies in cultivated forms and their wild relative as well as for comparative genetic mapping studies aimed at understanding patterns of genome rearrangement between *C. cardunculus* and related species.

1.5 Acknowledgements

We wish to thank:

- Loren H. Rieseberg, Steven J. Knapp and Zhao Lai (Compositae Genome Project) for founding the RAD tag and transcriptome sequencing within the U.S. National Science Foundation grants "Comparative genomics of phenotypic variation in the Compositae (DBI-0820451)"
- Giovanni Mauromicale and Rosario P. Mauro (Dipartimento di Scienze delle Produzioni Agrarie e Alimentari (DISPA) - Sez. Scienze Agronomiche, University of Catania) for the development and maintenance in field of the mapping progenies.

1.6 References

Acquadro A, Portis E, Lee D et al. (2005) Development and characterization of microsatellite markers in *Cynara cardunculus* L. Genome 48:217-225

Acquadro A, Lanteri S, Scaglione D et al. (2009) Genetic mapping and annotation of genomic microsatellites isolated from globe artichoke. Theor Appl Genet 118:1573-1587

Angelini LG, Ceccarini L, Di Nasso NNO, Bonari E (2009) Long-term evaluation of biomass production and quality of two cardoon (*Cynara cardunculus* L.) cultivars for energy use. Biomass Bioenerg 33:810-816

Bachlava E, Taylor CA, Tang S et al. (2012) SNP discovery and development of a high-density genotyping array for sunflower. PLoS one 7:e29814

Baird N, Etter P, Atwood T et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS one 3:e3376

Barbagallo RN, Chisari M, Spagna G et al. (2007) Caseinolytic activity expression in flowers of *Cynara cardunculus* L., Acta Hort 730:195-199

Blumenthal M, Goldberg A, Brinckmann J (2000) Artichoke leaf. eds. Herbal Medicine: Expanded Commission E Monographs, Integrative Medicine Communications Boston, MA, 10-210-12

Brown J, Rice-Evans C (1998) Luteolin-rich artichoke extract protects low density lipoprotein from oxidation in vitro. Free Rad Res 29:247-255

Bundy R, Walker A, Middleton R et al. (2008) Artichoke leaf extract (*Cynara scolymus*) reduces plasma cholesterol in otherwise healthy hypercholesterolemic adults: A randomized, double blind placebo controlled trial. Phytomedicine 15:668-675

Chatelet P, Stamigna C, Thomas G (2005) Early development from isolated microspores of *Cynara cardunculus* var. *scolymus* (L.) Fiori. Acta Hort 681:375-380

Chevreux B, Pfisterer T, Drescher B et al. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14:1147-1159

Chiapparino E, Lee D, Donini P (2004) Genotyping single nucleotide polymorphisms in barley by tetra-primer ARMS-PCR. Genome 47:414-420

Cho J, Kim A, Jung J et al. (2004) Cytotoxic and pro-apoptotic activities of cynaropicrin, a sesquiterpene lactone, on the viability of leukocyte cancer cell lines. Eur J Pharmacol 492:85-94

Comino C, Lanteri S, Portis E et al. (2007) Isolation and functional characterization of a cDNA coding a hydroxycinnamoyltransferase involved in phenylpropanoid biosynthesis in *Cynara cardunculus* L. BMC Plant Biol 7:14

Comino C, Hehn A, Moglia A et al. (2009) The isolation and mapping of a novel hydroxycinnamoyltransferase in the globe artichoke chlorogenic acid pathway. BMC Plant Biol 9:30

Cravotto G, Nano G, Binello A et al. (2005) Chemical and biological modification of cynaropicrin and grosheimin: a structure-bitterness relationship study. J Sc Food Agric 85:1757-1764

Davey JW, Hohenlohe PA, Etter PD et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12: 499-510

De Paolis A, Pignone D, Morgese A et al. (2008) Characterization and differential expression analysis of artichoke phenylalanine ammonia-lyase-coding sequences. Phys Plant 132:33-43

Emendorfer F, Emendorfer F, Bellato F et al. (2005) Antispasmodic activity of fractions and cynaropocrin from *Cynara scolymus* on guinea-pig ileum. Biol Pharm Bull 28:902-904

Foti S, Mauromicale G, Raccuia S et al. (1999) Possible alternative utilization of *Cynara* spp. I. Biomass, grain yield and chemical composition of grain. Ind Crop Prods 10:219-228

544 Gebhardt R (1997) Antioxidative and protective properties of extracts from leaves of the artichoke
545 (*Cynara scolymus* L) against hydroperoxide-induced oxidative stress in cultured rat
546 hepatocytes. Tox Appl Pharm 144:279-286
547 Gebhardt R (1998) Inhibition of cholesterol biosynthesis in primary cultured rat hepatocytes by
548 artichoke (*Cynara scolymus* L.) extracts. J Pharm Exp Therapy 286:1122-1128
549 Goremykin V, Salamini F, Velasco R, Viola R (2009) Mitochondrial DNA of *Vitis vinifera* and the
550 issue of rampant horizontal gene transfer. Mol Biol Evol 26:99-110
551 Grattapaglia D, Sederoff R. (1994) Genetic-linkage maps of *Eucalyptus grandis* and *Eucalyptus*
552 *urophylla* using a pseudo-testcross - mapping strategy and RAPD markers. Genetics 137:1121-
553 1137
554 Hoffmann L, Maury S, Martz F et al. (2003) Purification, cloning, and properties of an
555 acyltransferase controlling shikimate and quinate ester intermediates in phenylpropanoid
556 metabolism. J Biol Chem 278:95-103
557 Ierna A, Mauromicale G. (2010) *Cynara cardunculus* L. genotypes as a crop for energy purposes in
558 a Mediterranean environment. Biomass Bioenerg 34:754-760
559 Ierna A, Mauro RP, Mauromicale G (2012) Biomass, grain and energy yield in *Cynara cardunculus*
560 L. as affected by fertilization, genotype and harvest time. Biomass Bioenerg 36:404-410
561 Ishida K, Kojima R, Tsuboi M et al. (2010) Effects of artichoke leaf extract on acute gastric
562 mucosal injury in rats. Biol Pharm Bull 33:223-229
563 Jaillon O, Aury J, Noel B et al. (2007) The grapevine genome sequence suggests ancestral
564 hexaploidization in major angiosperm phyla. Nature 449:463-467
565 Jiang D, Ye Q, Wang F, Cao L (2010) The mining of *Citrus* EST-SNP and its application in cultivar
566 discrimination. Agricultural Sciences in China 9:79-190
567 Jimenez-Gomez J, Maloof J (2009) Sequence diversity in three tomato species: SNPs, markers, and
568 molecular evolution. BMC Plant Biol 9:85
569 Kraft K (1997) Artichoke leaf extract - Recent findings reflecting effects on lipid metabolism, liver
570 and gastrointestinal tracts. Phytomed 4:369-378
571 Lanteri S, Di Leo I, Ledda L et al. (2001) RAPD variation within and among populations of globe
572 artichoke cultivar 'Spinoso sardo'. Plant Breed 120:243-246
573 Lanteri S, Saba E, Cadinu M et al. (2004) Amplified fragment length polymorphism for genetic
574 diversity assessment in globe artichoke. Theor Appl Genet 108:1534-1544
575 Lanteri S, Acquadro A, Comino C et al. (2006) A first linkage map of globe artichoke (*Cynara*
576 *cardunculus* var. *scolymus* L.) based on AFLP, S-SAP, M-AFLP and microsatellite markers.
577 Theor Appl Genet 112:1532-1542
578 Lanteri S, Portis E, Acquadro A et al. (2011) Morphology and SSR fingerprinting of newly
579 developed *Cynara cardunculus* genotypes exploitable as ornamentals. Euphytica 184:311-321
580 Lattanzio V, Kroon PA, Linsalata V, Cardinali A (2009) Globe artichoke: A functional food and
581 source of nutraceutical ingredients. J Func Foods 1:131-144
582 Maccarone E, Fallico B, Fanella F et al. (1999) Possible alternative utilization of *Cynara* spp. II.
583 Chemical characterization of their grain oil. Ind Crops Prod 1:229-237
584 Martino V, Caffini N, Phillipson J et al. (1999) Identification and characterization of antimicrobial
585 components in leaf extracts of globe artichoke (*Cynara scolymus* L.). Acta Hort 501:111-114
586 Marushia RG, Holt JS (2006) The effects of habitat on dispersal patterns of an invasive thistle,
587 *Cynara cardunculus*. Biol Invas 8:577-593
588 Matsui T, Ogunwande IA et al. (2006) Anti-hyperglycemic potential of natural products. Mini Rev
589 Med Chem 6:349-356

590 Mauro R, Portis E, Acquadro A et al. (2008) Genetic diversity of globe artichoke landraces from
591 Sicilian small-holdings: implications for evolution and domestication of the species. Cons
592 Genet 10:431-440
593 McDougall B, King P, Wu B et al. (1998) Dicaffeoylquinic and dicaffeoyltartaric acids are selective
594 inhibitors of human immunodeficiency virus type 1 integrase. Antimicrob Agents Chemother
595 42:140-146
596 Menin B, Comino C, Moglia A et al. (2010) Identification and mapping of genes related to
597 caffeoylquinic acid synthesis in *Cynara cardunculus* L. Plant Sc 179:338-347
598 Menin B, Comino C, Portis E et al. (2012) Genetic mapping and characterization of the globe
599 artichoke (+)-germacrene A synthase gene, encoding the first dedicated enzyme for
600 biosynthesis of the bitter sesquiterpene lactone cynaropicrin. Plant Sc 190:1-8
601 Miccadei S, Di Venere D, Cardinali A et al. (2008) Antioxidative and apoptotic properties of
602 polyphenolic extracts from edible part of artichoke (*Cynara scolymus* L.) on cultured rat
603 hepatocytes and on human hepatoma cells. Nutr Cancer 60:276-283
604 Miller M, Dunham J, Amores A et al. (2007) Rapid and cost-effective polymorphism identification
605 and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17:240-248
606 Moglia A, Comino C, Portis E et al. (2009) Isolation and mapping of a C3′H gene (CYP98A49)
607 from globe artichoke, and its expression upon UV-C stress. Plant Cell Rep 28:963-974
608 Motzo R, Deidda M (1993) Anther and ovule culture in globe artichoke. J. Genet. Breed., 47:263-
609 266
610 Niggeweg R, Michael A, Martin C (2004) Engineering plants with increased levels of the
611 antioxidant chlorogenic acid. Nat Biotech 22:746-754
612 Palmer LE, Rabinowicz PD, O'Shaughnessy AL et al. (2003) Maize genome sequencing by
613 methylation filtrations. Science 302:2115-2117
614 Penalver R, Duranvila N, Lopez MM (1994) Characterization and pathogenicity of bacteria from
615 shoot tips of the globe artichoke (*Cynara scolymus*). Ann Appl Biol 125:501-513
616 Perez-Garcia F, Adzet T, Canigueral S (2000) Activity of artichoke leaf extract on reactive oxygen
617 species in human leukocytes. Free Radic Res 33:661-665
618 Pittlern M, Ernst E. (1998) Artichoke leaf extract for serum cholesterol reduction. Perfusion 11:338-
619 340
620 Portis E, Acquadro A, Comino C et al. (2005a) Genetic structure of island populations of wild
621 cardoon [*Cynara cardunculus* L. var. *sylvestris* (Lamk) Fiori] detected by AFLPs and SSRs.
622 Plant Sc 169:199-210
623 Portis E, Barchi L, Acquadro A et al. (2005b). Genetic diversity assessment in cultivated cardoon
624 by AFLP (amplified fragment length polymorphism) and microsatellite markers. Plant Breed
625 124:299-304
626 Portis E, Mauromicale G, Barchi L et al. (2005c). Population structure and genetic variation in
627 autochthonous globe artichoke germplasm from Sicily Island. Plant Sc 168:1591-1598
628 Portis E, Mauromicale G, Mauro R et al. (2009a) Construction of a reference molecular linkage
629 map of globe artichoke (*Cynara cardunculus* var. *scolymus*). Theor Appl Genet 120:59-70
630 Portis E, Acquadro A, Scaglione D et al. (2009b). Construction of an SSR-based linkage map for
631 *Cynara cardunculus*. Proceeding of the 8th Plant Genomics European Meeting, p 142
632 Portis E, Acquadro A, Longo A, Mauro R, Mauromicale G, Lanteri S. (2010) Potentiality of *Cynara*
633 *cardunculus* L. as energy crop. J Biotech 150:S165-S166
634 Portis E, Scaglione D, Acquadro A et al. (2012) Genetic mapping and identification of QTL for
635 earliness in the globe artichoke/cultivated cardoon complex. BMC Res Notes 5:252

636 Robba L, Carine M, Russell S, Raimondo F (2005) The monophyly and evolution of *Cynara* L.
637 (Asteraceae) *sensu lato*: evidence from the Internal Transcribed Spacer region of nrDNA. Plant
638 Syst Evol 253:53-64
639 Rondanelli M, Giacosa A, Orsini F et al. (2011) Appetite Control and Glycaemia Reduction in
640 Overweight Subjects treated with a Combination of Two Highly Standardized Extracts from
641 *Phaseolus vulgaris* and *Cynara scolymus*. Phytother Res 25:1275-1282
642 Rottenberg A, Zohary D, Nevo E (1996) Isozyme relationships between cultivated artichoke and the
643 wild relatives. Genetic Resources and Crop Evolution 43:59-62
644 Rottenberg A, Zohary D (2005) Wild genetic resources of cultivated artichoke. Acta Horti 681:307–
645 311
646 Scaglione D, Acquadro A, Portis E et al. (2009) Ontology and diversity of transcript-associated
647 microsatellites mined from a globe artichoke EST database. BMC Genomics 10:454
648 Scaglione D, Acquadro A, Portis E et al. (2012a). RAD tag sequencing as a source of SNP markers
649 in *Cynara cardunculus* L. BMC Genomics 13:3
650 Scaglione D, Lanteri S, Acquadro A et al. (2012b). Large-scale transcriptome characterization and
651 mass discovery of SNPs in globe artichoke and its related taxa. Plant Biotech J 10:956-969.
652 Schinor E, Salvador M, Ito I et al. (2004) Trypanocidal and antimicrobial activities of *Moquinia*
653 *kingii*. Phytomedicine 11.224-229
654 Schneider G, Thiele K 1974. Die Verteilung des Bitter-stoffes Cynaropicrin in der Artischocke.
655 Planta Medica 26:174-183
656 Sevcikova P, Glatz Z, Slanina J (2002) Analysis of artichoke (*Cynara cardunculus* L.) extract by
657 means of micellar electrokinetic capillary chromatography. Electrophoresis 23:249-252
658 Shao F, Merritt P, Bao Z et al. (2002) A *Yersinia* effector and a *Pseudomonas* avirulence protein
659 define a family of cysteine proteases functioning in bacterial pathogenesis. Cell 109:575-588
660 Shimoda H, Ninomiya K, Nishida N et al. (2003) Anti-hyperlipidemic Sesquiterpenes and new
661 sesquiterpene glycosides from the leaves of artichoke (*Cynara scolymus* L.): Structure
662 requirement and mode of action. Bioorg Med Chem Lett 3:223-228.
663 Slanina J, Taborska E, Bochorakova H et al. (2001) New and facile method of preparation of the
664 anti-HIV-1 agent, 1,3-dicaffeoylquinic acid. Tetrahedron Letters 42:3383-3385
665 Slanina J, Taborska E, Musil P (1993) Determination of cynarine in the decoctions of the artichoke
666 (*Cynara cardunculus* L.) by the HPLC method. Cesko-SloV Farm, 42:265-268
667 Sonnante G, D'Amore R, Blanco E et al. (2010) Novel hydroxycinnamoyl-coenzyme a quinate
668 transferase genes from artichoke are involved in the synthesis of chlorogenic acid. Plant
669 Physiol 153:1224-1238
670 Sonnante G, Pignone D, Hammer K (2007) The domestication of artichoke and cardoon: From
671 roman times to the genomic age. Ann Bot 100:1095-1100
672 Sonnante G, Gatto A, Morgese A et al. (2011) Genetic map of artichoke × wild cardoon: toward a
673 consensus map for *Cynara cardunculus*. Theor Appl Genet 123:1215-1229
674 Sousa MJ, Malcata FX (1996) Influence of pasteurization of milk and addition of starter cultures on
675 protein breakdown in ovine cheeses manufactured with extracts from flowers of *Cynara*
676 *cardunculus*. Food Chem 57:549-556
677 Stamigna C, Saccardo F, Pandozy G et al. (2005) In vitro mutagenesis of globe artichoke (cv.
678 Romanesco). Acta Hort 681:403-410
679 Timme R, Kuehl J, Boore J, Jansen R (2007) A comparative analysis of the *Lactuca* and *Helianthus*
680 (Asteraceae) plastid genomes: Identification of divergent regions and categorization of shared
681 repeats. Am J Bot 94:302-312
682 Velasco R, Zharkikh A, Troggio M et al. (2007) A High Quality Draft Consensus Sequence of the
683 Genome of a Heterozygous Grapevine Variety. PLoS one 2:e1326

684 Villegas R, Kojima M (1986) Purification and characterization of Hydroxycinnamoyl D-Glucose:
685 quinate hydroxycinnamoyl transferase in the root of sweet potato, *Ipomoea batatas* Lam. J Biol
686 Chem 261:8729-8733
687 Wagenbreth D (1996) Evaluation of artichoke cultivars for growing and pharmaceutical use. Beitr
688 Zuchtungsforsch 2:400-403
689 Wang M, Simon J, Aviles I et al. (2003) Analysis of antioxidative phenolic compounds in artichoke
690 (*Cynara scolymus* L.). J Agr Food Chem 51:601-608
691 Wiklund A (1992) The genus *Cynara* L. (Asteraceae, Cardueae). Bot J Linn Soc 109:75-123
692 Yasukawa K, Matsubara H, Sano Y (2010) Inhibitory effect of the flowers of artichoke (*Cynara*
693 *cardunculus*) on TPA-induced inflammation and tumor promotion in two-stage carcinogenesis
694 in mouse skin. J Nat Med 64:388-391
695 Ye S, Dhillon S, Ke X et al. (2001) An efficient procedure for genotyping single nucleotide
696 polymorphisms. Nucleic Acids Res 29:e88-8

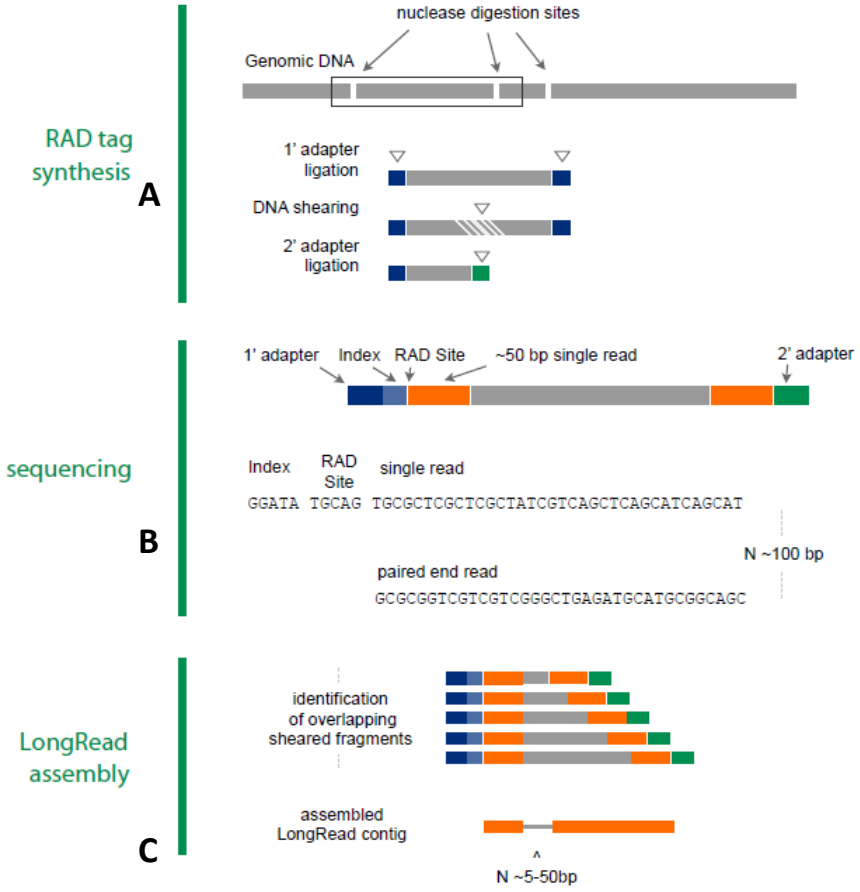
BOX 1

RADseq (Restriction-site Associated DNA sequencing)

An efficient approach for SNP discovery, is RAD "Restriction-site Associated DNA" (Miller et al. 2007), coupled with NGS technologies (Baird et al. 2008), which has been recently termed as RADseq (Davey et al. 2011). At least 20 papers have been recently published in both animals (snails, moths and salmon, sturgeon, butterflies, beetles and worms) and plants (ryegrass, oaks, lolium, eggplant and globe artichoke). A detail review is available at the wiki RAD-sequencing page (University of Edinburgh; https://www.wiki.ed.ac.uk/display/RADSequencing).

The strategy requires the enzymatic digestion of a genome with at least one restriction enzyme and the sequencing of the resulting fragments through an Illumina Genome Analyzer. The fragments from one sample are ligated to a modified Illumina adapter containing a unique identifying sequence (Molecular IDentifier, or MID). A list of the available primers can be found at the above-cited wiki RAD-sequencing section. The fragments from many samples (e.g. a mapping population) can consequently be pooled together and sequenced on a single lane. The resulting reads can be segregated using the MID present at the start of each read. By sequencing simultaneously all the individuals of a population of interest, and by comparing the tags, thousand of SNPs at different genetic loci can be identified in a single experiment.

The protocol is depicted in the figure reported below. **A)** Genomic DNA is digested with a restriction enzyme and a barcoded P1 adapter is ligated to the fragments. The P1 adapter contains a forward amplification primer site, an Illumina sequencing primer site, and a barcode for sample identification. Adapter-ligated fragments are pooled (if multiplexing), sheared, size-selected (e.g. 300-800 bp) and ligated to a second adapter (P2). The P2 adapter is a divergent "Y" adapter, containing the reverse complement of the P2 reverse-amplification primer site, preventing amplification of genomic fragments lacking a P1 adapter. **B)** The samples are analysed on an Illumina Genome Analyzer IIx following the paired ends (2x 54 bp, or more) genomic DNA sequencing protocol. The generated sequences are then sorted according to their multiplex identifier tag (barcode). **C)** The sequences are *de novo* assembled using a bioinformatics DNA assembler (e.g.: Velvet). Assembled LongRead® contigs can be generated by a set of algorithms developed at Floragenex Inc. (Oregon, USA).



GO ID	Term	N° of loci	EST-SSR loci
GO:0050896	response to stimulus	17	CyEM-008, -030, -42, -054, -057, -070, -072, -093, -120, -135, -145, -150, -152, -218, -229, -259, -266
GO:0009628	response to abiotic stress	13	CyEM -008, -030, -054, -070, -093, -120, -135, -145, -150, -152, -218, -229, -259
GO:0042221	response to chemical stimulus	4	CyEM -093, -218, -229, -266
GO:0006950	response to abiotic stress	15	CyEM -008, -030, -054, -057, -070, -072, -093, -120, -135, -145, -150, -152, -229, -259, -266
GO:0009266	response to temperature stress	5	CyEM -008, -054, -093, -145, -150
GO:0006970	response to osmotic stress	8	CyEM -030, -070, -093, -120, -135, -152, -229, -259
GO:0010033	response to organic substance	3	CyEM -093, -229, -266
GO:0009409	response to cold stress	5	CyEM -008, -054, -093, -145, -150
GO:0009651	response to salt stress	8	CyEM -030, -070, -093, -120, -135, -152, -229, -259

Table 1: CyEM (*Cynara* Expressed Microsatellites) markers with Gene Ontology annotation for stimuli response-related terms.

#	Genotype	*C. carduncul*us taxon	Sequencing results			Assembly results	
			Raw reads	Total (Mbp)	Mean length (bp)	Contigs	Mean length/N50 (bp)
1	'Romanesco C3'	var. *scolymus*	0.43 M	184	421	37,622	834/723.8
2	'Altilis 41'	var. *altilis*	0.61 M	246	402	40,130	761/699.9
3	'Creta 4'	var. *sylvestris*	0.69 M	263	377	42,837	772/688.5
Total			**1.74 M**	**693**	**392**	**38,726***	**951/844.3***

Table 2. 454-derived sequencing and assembly. The output statistics were calculated following the removal of contaminating and adaptor sequences. Data are intended after quality filtering and sequence clipping. *Asterisks indicate results obtained by merging the three independent assemblies (see Figure 3).

#	Genotype	*C. cardunculus* taxon	Raw reads	First mates (Mbp)	Paired mates (Mbp)
4	'Romanesco Zorzi'	var. *scolymus*	6.6M	458	408
5	'Violetto di Chioggia'	var. *scolymus*	6.6M	470	420
6	'Violetto Pugliese'	var. *scolymus*	5.2M	367	331
7	'Spinoso Sardo'	var. *scolymus*	6.7M	474	424
8	'Imperial Star'	var. *scolymus*	6.4M	459	415
9	'Blanco de Peralta'	var. *altilis*	4.8M	340	305
10	'Gobbo di Nizza'	var. *altilis*	5.6M	380	341
11	'Sylvestris_LOT23'	var. *sylvestris*	4.6M	322	287
Total			**46.5M**	**3,271**	**2,931**

Table 3. GAIIx (Illumina)-derived sequencing. A total of 46.5 M raw reads were generated in two GAIIx lanes and 6.7 Gbp were retained after removing adaptor and contaminating sequences. The windowed quality clipping routine produced a final dataset of 6.2 Gbp. A higher number of bases were obtained for single ends, because 84 sequencing cycles were used instead of the 76 used for the paired ends.

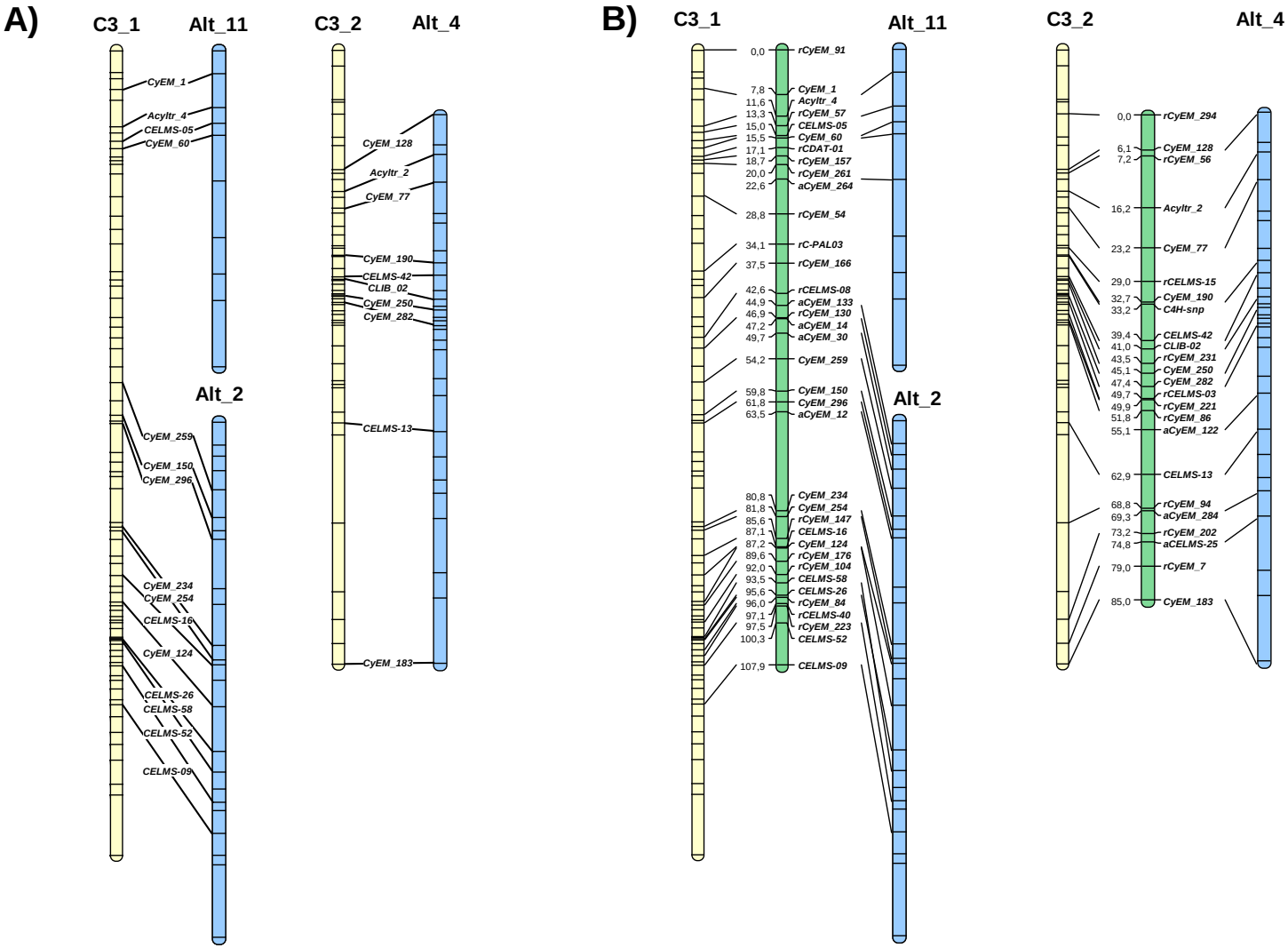


Figure 1: Examples of alignment and consensus LG construction. Alignment of the ‘Romanesco C3’ (yellow) and the ‘Altilis 41’ (blue) LGs based on common markers (A). SSR-based consensus LGs (green) construction (B). 'r-' and 'a-' indicate markers segregating only in, respectively, ‘Romanesco C3’ and ‘Altilis 41’. Marker nomenclature is the one reported in Portis et al. (2009) and Scaglione et al. (2009).

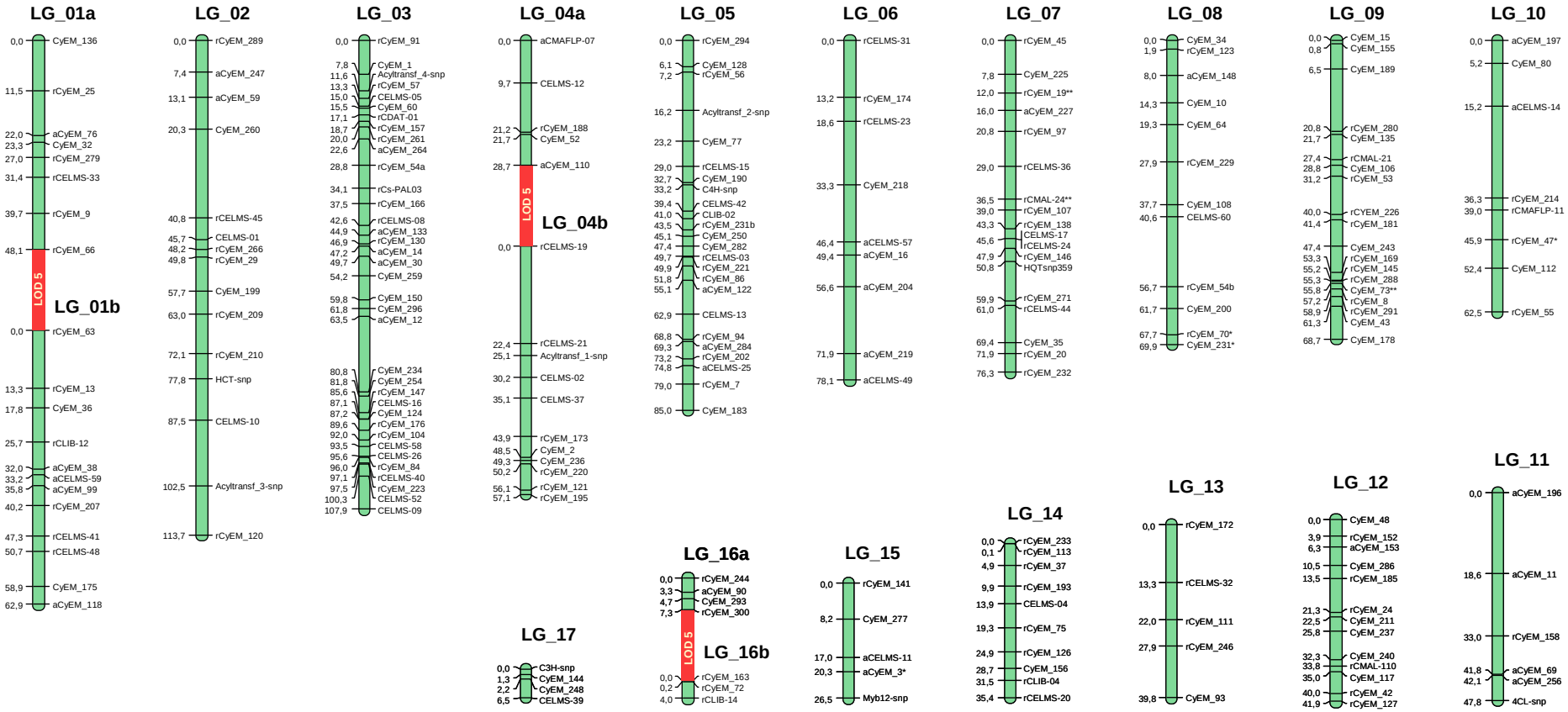


Figure 2: The SSR-based consensus map of *C. cardunculus*. Marker names are shown to the right of each LG, with map distances (in cM) to the left. 'r-' and 'a-' indicate markers segregating only in, respectively, 'Romanesco C3' and 'Altilis 41'. Segments shaded in red indicate where a pair of LGs has merged as a result of reducing the stringency to LOD 5. Marker nomenclature is the one reported in Portis et al. (2009) and Scaglione et al. (2009).

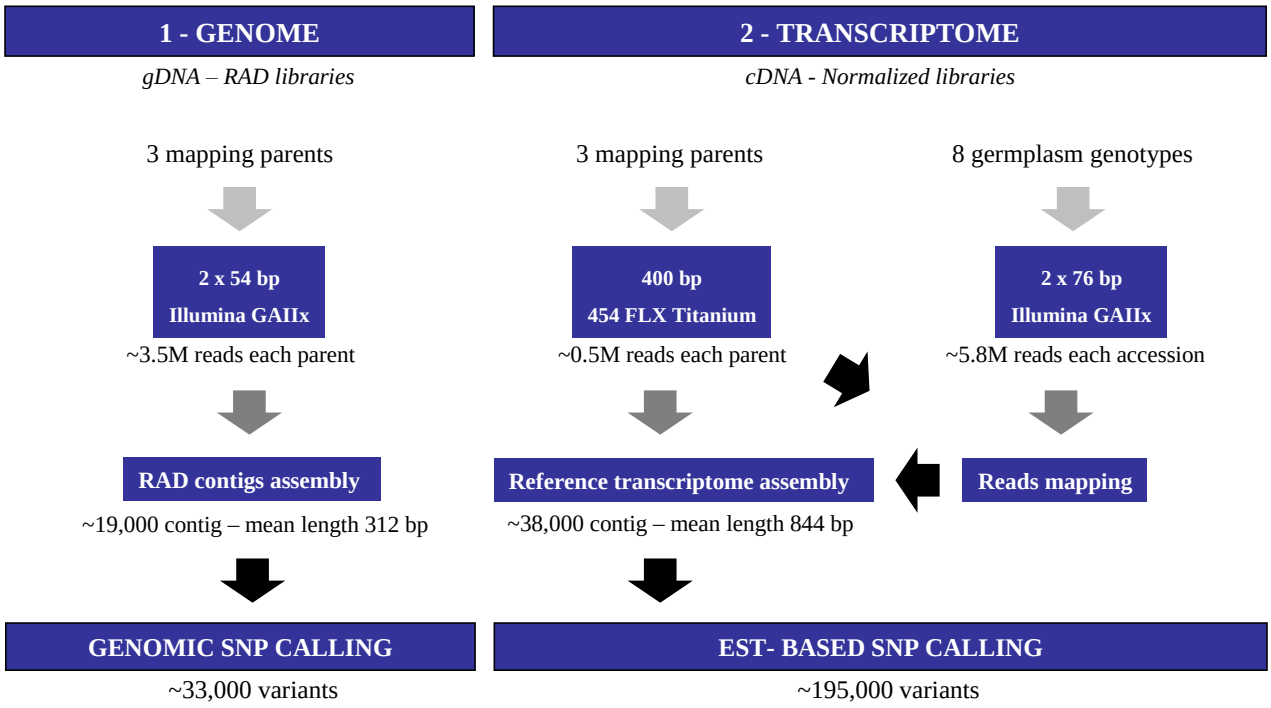


Figure 3. SNP mining workflow in *Cynara cardunculus*.

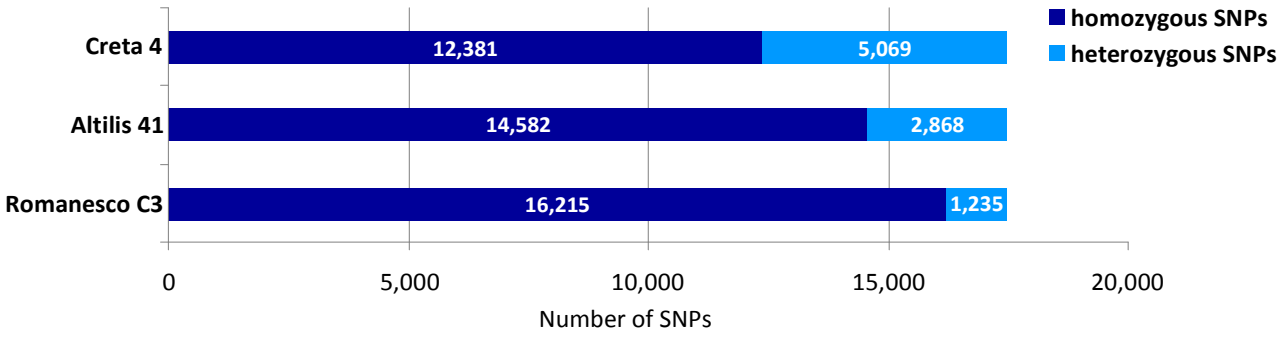


Figure 4: Proportion of heterozygous SNPs across the three mapping parents.

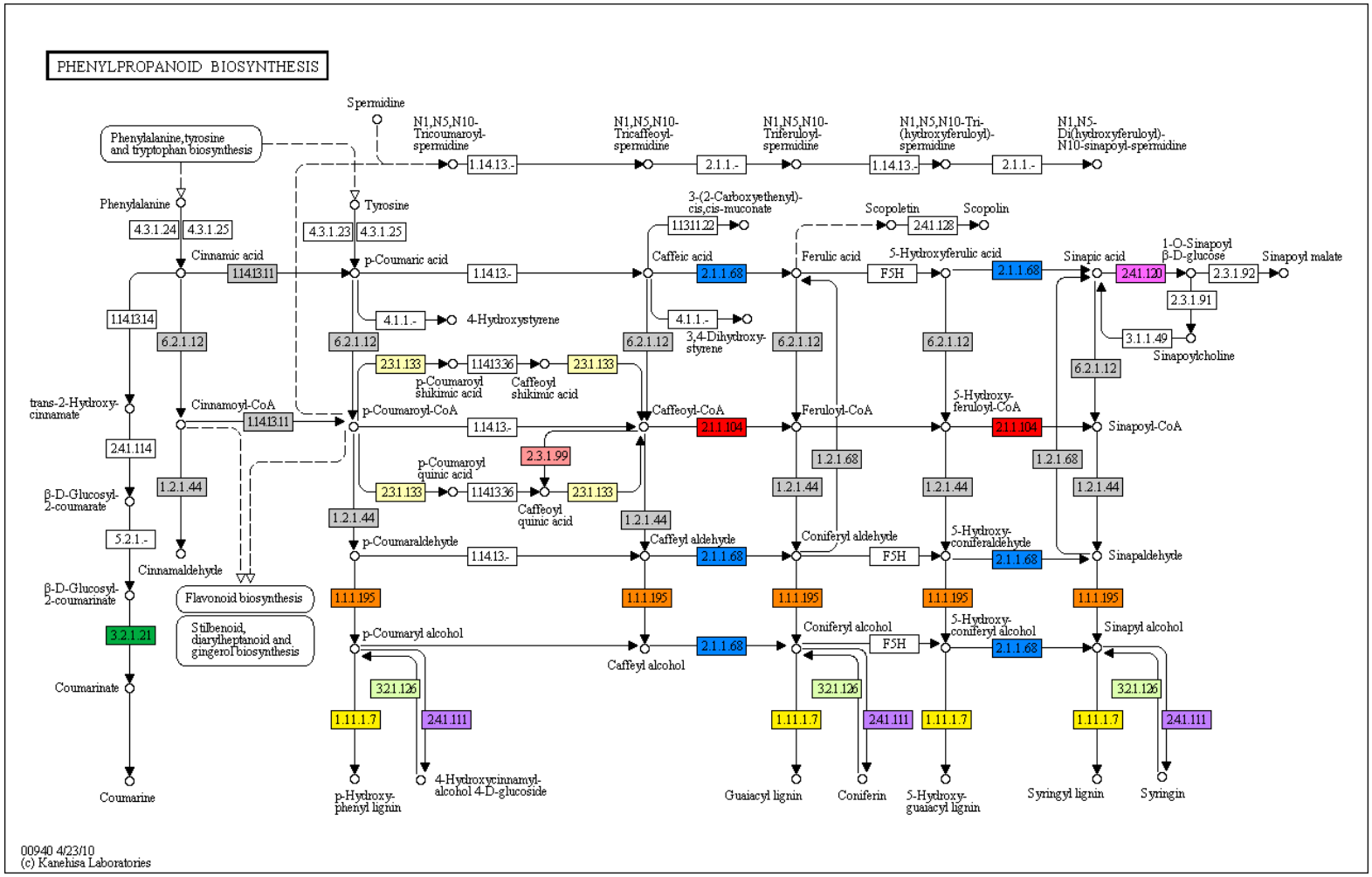


Figure 5: *C. cardunculus* phenylpropanoid enzymatic activities (coloured) mapped on KEGG (map: 00940).

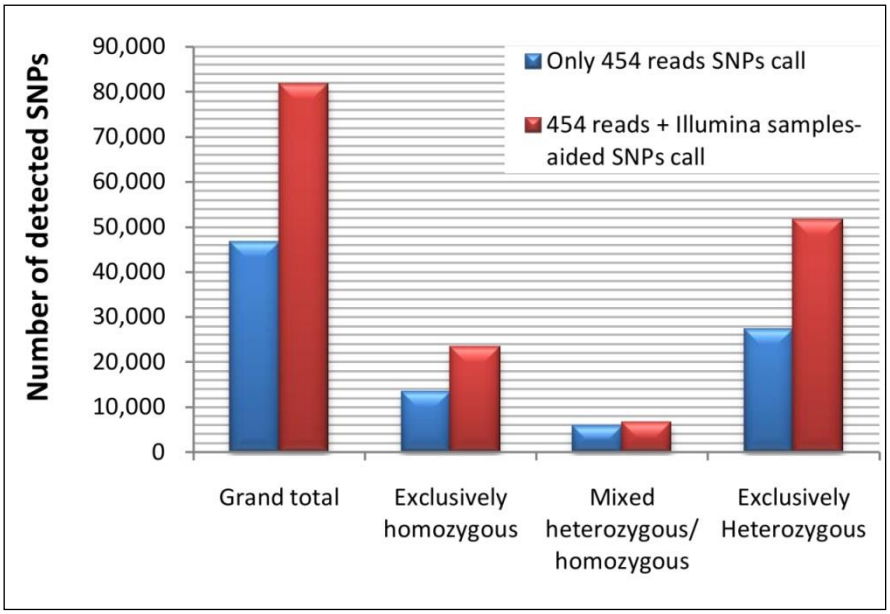


Figure 6: Combined calling of SNPs. The number of calls based solely on the 454-derived reads is shown in blue, and the combined SNP discovery based on both the 454- and the Illumina-based sequence in red. "Exclusively homozygous" and "exclusively heterozygous" refer to allelic variants present in only one of the three 454-sequenced libraries.

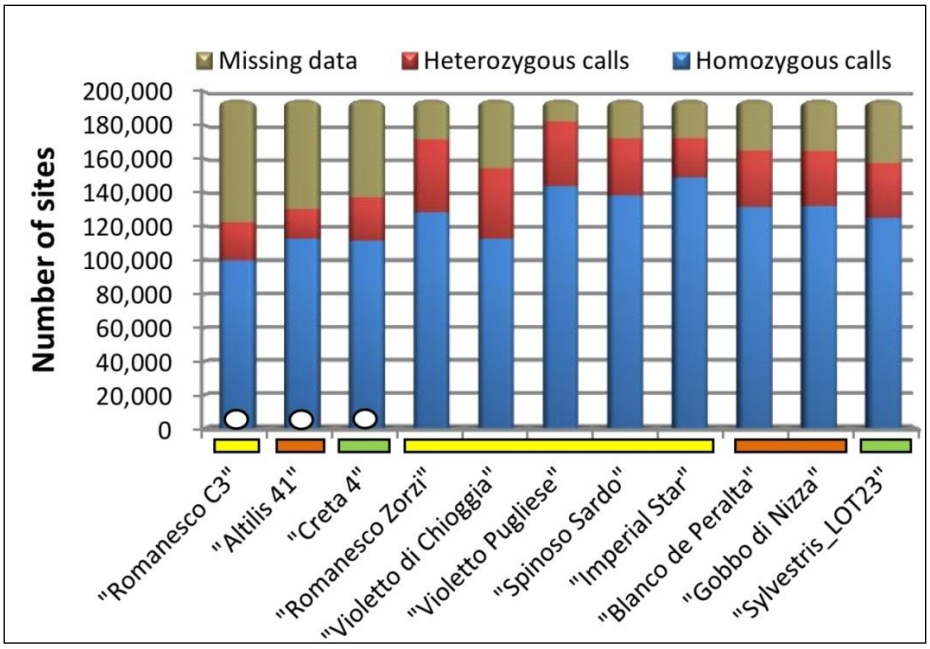


Figure 7: The allelic state at SNP loci. Bars indicate the total number of SNP loci in the homozygous or heterozygous state (or missing) for each accession. Each bar's colour identifies the *C. cardunculus* taxa (green = *sylvestris*, orange = *altilis*, yellow = *scolymus*). White dots identify the three accessions sequenced using 454 technology.